

## ORIGINAL ARTICLE

# Standardisation of the Test Your Memory and evaluation of their concordance with the outcome of the psychometric examination<sup>☆</sup>



J. Ferrero-Arias\*, M.A. Turrión-Rojo

Centro Médico Orense, Madrid, Spain

Received 3 December 2014; accepted 24 March 2015  
Available online 6 May 2016

### KEYWORDS

Test Your Memory;  
Mild cognitive impairment;  
Psychometrics;  
Dementia;  
Alzheimer disease;  
Cognitive screening test

### Abstract

**Objective:** To explore the relationship between scores on the Test Your Memory (TYM) battery and findings from a more exhaustive neurocognitive assessment.

**Methods:** The TYM and fourteen psychometric tests were administered to 84 subjects aged 50 or older who attended an outpatient neurology clinic due to cognitive symptoms. Each patient's cognitive state was determined independently from his/her score on the TYM (CDR 0,  $n = 25$ ; CDR 0.5,  $n = 45$ ; CDR 1,  $n = 14$ ). We analysed concurrent validity of TYM scores and results from the psychometric tests, as well as the degree of concordance between the two types of measurement, by contrasting normalised data from each instrument.

**Results:** Although the intraclass correlation coefficient was 0.67 (confidence interval 95%, 0.53–0.77), analysis of the Bland–Altman plot and the curve on the survival-agreement plot (Luiz et al. method) demonstrates that the individual distances between the two methods exhibit excessive dispersion from a clinical viewpoint. TYM-based predictions of the mean z-score on psychometric tests differed substantially from real results in 30% of the subjects. Concordance of 95% can only be achieved by accepting absolute inter-instrument differences of up to 0.87 as identical values. Furthermore, the TYM underestimates cognitive performance for low values and overestimates it for high values.

**Conclusions:** The TYM is a cognitive screening test which should not be used to predict results on psychometric tests or to detect cognitive changes in clinical trials.

© 2014 Sociedad Española de Neurología. Published by Elsevier España, S.L.U. All rights reserved.

<sup>☆</sup> Please cite this article as: Ferrero-Arias J, Turrión-Rojo MA. Normalización del Test Your Memory y evaluación de su concordancia con los resultados del examen psicométrico. Neurología. 2016;31:239–246.

\* Corresponding author.

E-mail address: [jferreroarias@gmail.com](mailto:jferreroarias@gmail.com) (J. Ferrero-Arias).

**PALABRAS CLAVE**

Test Your Memory;  
Deterioro cognitivo  
ligero;  
Psicometría;  
Demencia;  
Enfermedad de  
Alzheimer;  
Prueba cognitiva de  
cribado

## Normalización del Test Your Memory y evaluación de su concordancia con los resultados del examen psicométrico

**Resumen**

**Objetivo:** Explorar la relación entre las puntuaciones del Test Your Memory (TYM) y los hallazgos observados en una valoración neurocognitiva más extensa.

**Métodos:** Se administraron el TYM y 14 pruebas psicométricas a 84 sujetos de 50 o más años que acudieron a una consulta neurológica ambulatoria con sintomatología de estirpe cognitiva. Se determinó su estado cognitivo con independencia del resultado del TYM (CDR 0: 25, CDR 0.5: 45, CDR 1: 14). Se analizaron la validez concurrente del TYM con los hallazgos de la psicometría, así como el grado de concordancia entre ambos tipos de medida, confrontando datos normalizados y tipificados de los instrumentos.

**Resultados:** Aunque el coeficiente de correlación intraclase fue de 0,67 (intervalo de confianza del 95%, 0,53-0,77), el análisis mediante el gráfico de Bland-Altman y la curva de supervivencia-acuerdo de Luiz et al. demuestran que las distancias individuales entre ambos métodos tienen una dispersión excesiva desde el punto de vista clínico. Los valores predichos por el TYM de la puntuación z media de las pruebas psicométricas discrepan sustancialmente en el 30% de los sujetos y solo aceptando como valores idénticos diferencias absolutas de hasta 0,87 puntos z entre ambas medidas se alcanza un 95% de concordancia. Además, el TYM subestima los rendimientos cognitivos cuando sus valores son bajos y los sobreestima si son elevados.

**Conclusiones:** El TYM es una prueba cognitiva de cribado que no debe utilizarse para predecir el resultado de la psicometría o determinar la ocurrencia de cambio cognitivo en ensayos clínicos.

© 2014 Sociedad Española de Neurología. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

**Introduction**

According to some studies, scores on certain cognitive screening tests, such as the Mini-Mental State Examination,<sup>1</sup> are acceptable predictors of the results of psychometric tests and representative of a patient's general cognitive state.<sup>2,3</sup> However, a good cognitive screening test may have a high detection capacity while inaccurately reflecting a patient's cognitive state. We have used the results of a recently validated Spanish-language version<sup>4</sup> of the Test Your Memory (TYM) test to determine whether using screening tools in neurocognitive evaluations is appropriate. Our study presents the results of the normalisation of the TYM on the original patient sample. The norms obtained were then applied to a new patient sample. The TYM scores in this sample were compared to the results of a more comprehensive neurocognitive evaluation that was conducted simultaneously.

**Subjects, material, and methods****Normalisation and classification procedure**

We used the original patient sample for normalisation and classification of the TYM battery.<sup>4</sup> This sample included 739 cognitively normal individuals aged 50 or older, with at least 4 years of schooling and a Clinical Dementia Rating (CDR)<sup>5,6</sup> score of 0.

Our methodology was similar to that applied in other normalisation studies.<sup>7,8</sup> Sex did not affect TYM scores in the

validation study<sup>4</sup>; this variable was therefore not considered in normalisation. Individuals were grouped into 5-year bands and raw TYM scores were sorted by percentile for each age group. Values were subsequently converted to scaled scores (mean  $\pm$  SD:  $10 \pm 3$ ) so as to normalise the distribution. We conducted a linear regression analysis to correct for education level, which is the variable affecting raw TYM scores the most. The dependent variable was scaled TYM scores (adjusted by age group), whereas the predictor variable was education level, which comprised 4 different levels (0: 4-8 years of schooling; 1: 9-12 years; 2: 13-15 years; and 3: more than 15 years). We then adjusted scaled scores by education level using the following equation:

$$TYM_{sc} = TYM_s - \beta \times \text{ed-level}$$

$TYM_{sc}$  represents scaled TYM scores corrected by years of schooling,  $TYM_s$  represents scaled TYM scores corresponding to the individual's age group,  $\beta$  is the regression coefficient obtained from the linear regression analysis, and 'ed-level' is the individual's education level (as stated previously).

Scaled scores, which were adjusted for age and education level, were converted to z scores ( $TYM_z = [TYM_{sc} - 10]/3$ ) to make it easier to compare TYM scores with z scores from other psychometric tests.

**Comparison group**

This group included patients aged 50 or older with at least 4 years of schooling who attended the neurology clinic of

one of the authors (JFA), and who in one day underwent a cognitive examination and were administered the TYM battery and other standardised neurocognitive tests. We ruled out patients with focal lesions (for example, cerebral infarcts or tumours) and only included those with CDR scores < 2 (CDR 0: 25 patients; CDR 0.5: 45; CDR 1: 14). CDR scores were calculated with a Spanish-language structured questionnaire available from <http://alzheimer.wustl.edu/cdr/PDFs/Translations/Spanish%20Spain.pdf>.

Patients in the comparison group were not included in the patient sample used for normalisation. For each patient, we calculated the TYMz corresponding to their raw TYM scores, adjusting for age group and education level.

## Neurocognitive evaluation

All patients in the comparison group took the TYM, forward and backward digit span,<sup>9</sup> semantic verbal fluency (animals/min) and phonemic verbal fluency (words beginning with 'p'/min),<sup>10</sup> Free and Cued Selective Reminding Test (free and total recall, delayed free and total recall),<sup>11</sup> Boston naming test,<sup>12</sup> parts A and B of the Trail Making Test,<sup>9</sup> and Rey–Osterrieth Complex Figure Test (time to copy, copy accuracy, and immediate recall [3 minutes]).<sup>11</sup> Test results were expressed as z scores. We calculated the mean z score of all the applied neurocognitive tests (mean-z), except for TYM, as an indicator of general cognitive performance. Some studies have shown that a set of neurocognitive tests similar to that in our study satisfactorily evaluates general cognitive performance.<sup>13,14</sup>

## Statistical analysis

Concurrent validity was tested by calculating the intraclass correlation coefficient between TYMz and mean-z. We then calculated the mean-z predicted by TYMz (Pmean-z) by using a linear regression equation where the criterion variable was mean-z and the predictor variable was TYMz. Agreement and the limits of agreement were evaluated using a Bland–Altman plot.<sup>15</sup> Using the Bland–Altman plot simplified the assessment of bias. Once Pmean-z was calculated, we determined whether the limits of agreement with mean-z were acceptable from a clinical viewpoint. This latter aspect was also evaluated using the 'survival-agreement plot' designed by Luiz et al.<sup>16</sup> We estimated the reliable change index (RCI) of the 2 measurement tools with the formula described by Jacobson and Truax<sup>17</sup> using the standard error of the measurement (SEM) and calculating 95% CI ( $\pm \text{SEM} \times \sqrt{2} \times 1.96$ ):

$$\text{SEM} = \text{SD} \times \sqrt{(1 - r)},$$

where  $r$  is the test–retest reliability coefficient of each test.

In the case of mean-z, the test–retest reliability coefficient used for calculating the RCI was 0.84 (this was obtained using an independent sample of 49 control patients). The internal consistency of the set of 14 tests (which integrate a 'virtual scale') was calculated using the same independent

sample of 49 patients and was shown to be sufficient, with a Cronbach's alpha value of 0.85 (95% lower CI limit of 0.80). On the other hand, the RCI of TYMz was calculated with the normative subsample used for test–retest (42 individuals).<sup>4</sup> Minimum detectable change was calculated using the RCI of both mean-z and TYMz (RCIm).

This analysis assumed an alpha risk of 0.05 in a bilateral contrast. When necessary, estimates are provided with their 95% CI. Statistical analysis was conducted using Stata/IC™ 12.1 and MedCal™ 14.10.2.

## Results

### Normalisation and classification

Table 1 summarises the sample's sociodemographic characteristics and scores on each cognitive test, broken down by CDR score; cognitive tests are arranged by cognitive domain. Table 2 shows raw TYM scores converted to scaled scores, broken down by age group. Age adjustment was subsequently performed using the equation  $\text{TYM}_{sc} = \text{TYM}_s - 0.811 \times \text{ed-level}$ , obtained from the linear regression analysis. RCI of TYMz was 0.57 and RCI of mean-z was 0.61; therefore, RCIm was  $\pm 0.59$  z points (95% CI limits). Supplementary material to this article includes an Excel application that calculates scaled and z scores using age, years of education, and raw TYM scores (Appendix).

### Comparison sample

Table 3 shows Pearson correlation coefficients between TYMz and z scores for each test. The tests most strongly correlated with TYM are the ones evaluating verbal memory and language, whereas executive function tests displayed a weaker, though still statistically significant, correlation with TYM. The intraclass correlation coefficient between TYMz and mean-z was 0.67 (95% CI, 0.53–0.77) ( $R^2 = 0.45$ ). Linear regression mean-z–TYMz produced the following equation:  $\text{mean-z} = -0.1754 + 0.3773 \times \text{TYMz}$ . The Bland–Altman plot, which was constructed using the data corresponding to mean-z and TYMz (Fig. 1), shows that the difference between the 2 measures was greater than the 95% CI for a mean difference of 0, and the limits of agreement are clearly more than  $\pm 1$  SD. Fig. 1 also shows that TYMz overestimates general cognitive performance in neurocognitive tests for high values and underestimates it for low values, indicating that calibration is imprecise for extreme values. In the Bland–Altman plot created with mean-z and Pmean-z (Fig. 2), the differences were centered within the CI limits for a mean difference of 0, and the limits of agreement narrowed ( $\pm 0.87$  z points). Considering a difference between z values higher than 0.59 z points (RCIm) as clinically significant, the percentage of disagreement reaches 30% of the cases, as can be seen in the survival-agreement plot (Fig. 3). To reach 95% agreement, there must be a difference of at least 0.87 z points between both measurements, that is, 0.28 z points more than the RCIm.

**Table 1** Sociodemographic characteristics and scores on each cognitive test, broken down by CDR score.

	CDR 0	CDR 0.5	CDR 1
<i>n</i>	25	45	14
Age (mean ± SD)	71.0 ± 10.1	71.1 ± 9.2	73.3 ± 6.4
Sex (% women)	67.0	69.2	30.0
Years of schooling (mean ± SD)	14.1 ± 4.5	12.2 ± 4.8	15.1 ± 4.4
<b>Cognitive domain</b>			
<i>Attention</i>			
Forward digit span (mean ± SD)	0.30 ± 0.80	0.10 ± 0.74	-0.43 ± 1.55
Backward digit span	0.49 ± 0.74	0.15 ± 0.71	0.10 ± 0.96
<i>Executive function</i>			
Trail Making Test, part A	-0.05 ± 0.95	-0.71 ± 1.00	-1.13 ± 1.25
Trail Making Test, part B	-0.16 ± 0.99	-0.89 ± 0.88	-1.78 ± 1.25
Rey-Osterrieth Complex Figure Test (time to copy)	-0.31 ± 0.94	-0.16 ± 0.79	-0.96 ± 0.56
Phonemic verbal fluency ('p'/min)	0.34 ± 0.81	-0.21 ± 0.92	-0.57 ± 1.08
<i>Language</i>			
Semantic verbal fluency (animals/min)	0.02 ± 0.92	-0.81 ± 0.89	-2.15 ± 0.91
Boston naming test	0.57 ± 0.83	-0.09 ± 1.13	-0.99 ± 0.83
<i>Memory</i>			
FCSRT free total recall (1 + 2 + 3)	-0.21 ± 0.73	-1.05 ± 0.77	-2.22 ± 0.91
FCSRT total recall (1 + 2 + 3)	-0.16 ± 0.87	-0.93 ± 0.93	-2.11 ± 1.03
FCSRT delayed free recall	-0.21 ± 0.71	-1.73 ± 1.13	-2.43 ± 1.17
FCSRT delayed total recall	-0.35 ± 0.78	-1.28 ± 1.36	-2.30 ± 0.93
<i>Visuospatial function</i>			
Rey-Osterrieth Complex Figure Test (copy accuracy)	0.04 ± 1.05	-0.30 ± 0.93	-0.39 ± 1.37
Rey-Osterrieth Complex Figure Test (immediate recall)	-0.40 ± 0.68	-1.21 ± 0.70	-2.13 ± 1.01
Mean-z	0.03 ± 0.42	-0.65 ± 0.36	-1.35 ± 0.56
TYMz	-0.02 ± 0.78	-1.82 ± 0.86	-2.00 ± 0.90

CDR: Clinical Dementia Rating; mean-z: mean z score of all the applied neurocognitive tests except for TYM; FCSRT: Free and Cued Selective Reminding Test.

Test scores are expressed as mean z score ± SD.

**Table 2** Conversion from raw TYM scores to scaled scores, broken down by age group.

Scaled scores	Percentiles	Age groups <sup>a</sup>							
		50–54	55–59	60–64	65–69	70–74	75–79	80–84	>85
2	<1	<39	<36	<36	<35	<32	<32	<31	<30
3	1	–	–	–	35	32–35	32	31	–
4	2	–	36	36	36	36	33–34	32–35	–
5	3–5	39–40	37–38	37–38	37–38	37–38	35–37	36	30
6	6–10	41–42	39–40	39–40	39	39	38	37–38	31–33
7	11–18	43	41	41–42	40–41	40–41	39–40	39	34–35
8	19–28	44	42–43	43	42	42	41	40–41	36–38
9	29–40	45	44	44	43–44	43	42	42	39–40
10	41–59	46–47	45–46	45–46	45–46	44	43–44	43–44	41–42
11	60–71	48	47	47	47	45–46	45	45–46	43
12	72–81	–	–	48	48	47	46	47	44
13	82–89	49	48	49	49	48	47–48	48	45
14	90–94	–	49	–	–	49	49	49	46–47
15	95–97	–	–	–	–	–	–	–	–
16	98	–	–	–	–	–	–	–	48
17	99	–	–	–	–	–	–	–	49
18	>99	50	50	50	50	50	50	50	50
Sample	739	40	50	116	129	122	121	104	57

<sup>a</sup> Data obtained from the control group in the study by Ferrero-Arias and Turrión-Rojo.<sup>4</sup>

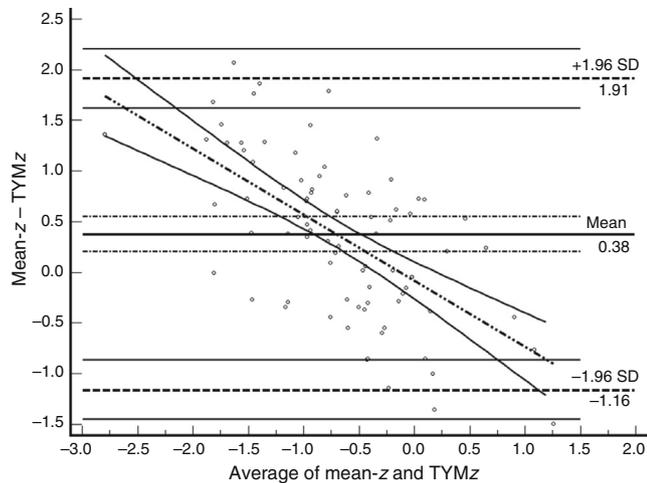
**Table 3** Correlation between psychometric tests and TYM.

	CC
FCSRT free total recall (1 + 2 + 3)	0.685****
FCSRT total recall (1 + 2 + 3)	0.684****
FCSRT delayed total recall	0.639****
FCSRT delayed free recall	0.593****
Boston naming test	0.548****
Rey–Osterrieth Complex Figure Test (immediate recall [3 min])	0.509****
Semantic verbal fluency (animals/min)	0.464****
Forward digit span	0.416****
Rey–Osterrieth Complex Figure Test (copy accuracy)	0.412****
Trail Making Test part B (time)	0.405****
Backward digit span	0.372****
Phonemic fluency (words beginning with 'p')	0.314**
Trail Making Test part A (time)	0.285**
Rey–Osterrieth Complex Figure Test (time to copy)	0.265*

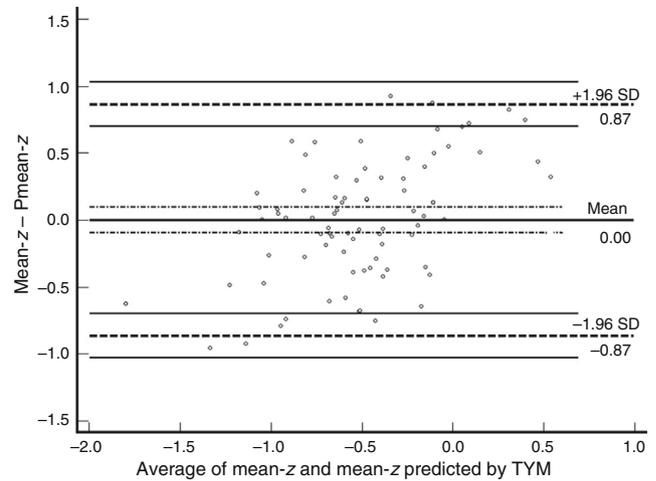
n: 84 patients.

CC: Pearson correlation coefficient between psychometric tests and TYM (z scores); FCSRT: Free and Cued Selective Reminding Test.

- \* P < .05.
- \*\* P < .01.
- \*\*\* P < .001.
- \*\*\*\* P < .0001.



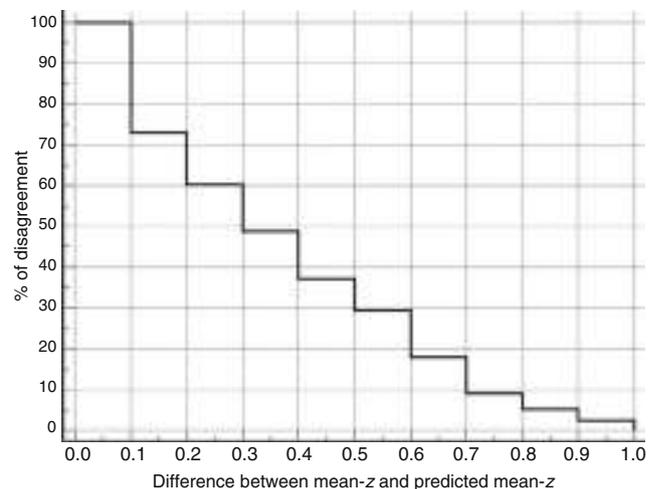
**Figure 1** Bland–Altman plot using mean-z and TYMz. The x-axis represents the average of mean-z and TYMz while the y-axis represents the differences between the 2 paired values of those parameters. We drew the line of equality, the lines for the limits of agreement (mean difference  $\pm$  1.96SD of the differences, with their respective 95% CI), and the regression line of the differences. The slope clearly indicates that TYM underestimates cognitive performance in neurocognitive tests for low values (positive differences) and overestimates it for high values (negative differences).



**Figure 2** Bland–Altman plot using mean-z and Pmean-z, which is calculated using the formula  $\text{mean-z} = -0.1754 + 0.3773 \times \text{TYMz}$ . Note that the mean difference is centered around 0 and the limits of agreement are narrower.

### Discussion

TYM values correlated well with language and verbal memory tests, and to a much lesser extent with executive function tests; as a result, we conclude that TYM assesses language and memory function more accurately than other cognitive domains. This tool is better suited for screening for Alzheimer disease (AD) or other types of dementia initially affecting language than it is for screening for the behavioural variant of frontotemporal dementia or certain forms of AD with an unusual presentation (posterior cortical atrophy). The normalisation of TYM carried out in this study defines the most suitable cut-off points for each age group and education level more accurately than the general



**Figure 3** Survival-agreement plot showing the percentage of disagreement between mean-z and Pmean-z; 95% of agreement is achieved only when a difference greater than 0.87 z points is found between z scores.

rule provided in the initial validation study of our Spanish-language version of the test.<sup>4</sup> However, we feel that the normative data from this study should be applied with caution since the TYM results do not seem to agree with those from a more exhaustive neurocognitive assessment.

The idea that brief cognitive tests can predict the results of and thus replace a more exhaustive neurocognitive assessment is not new.<sup>2,3,18–22</sup> This suggests that both types of assessment somehow reflect the same dimension: general cognitive performance as a reflection of a patient's global cognitive state. The literature has adopted this concept and uses these expressions literally.<sup>14</sup> We feel that the idea of 'general cognition' is rather elusive, but we accept it as an operative concept. In addition to their utility as detection or screening tools, brief cognitive tests are also used to assess impairment and detect and accurately reflect changes in patients' cognitive function. However, these tests have obvious disadvantages, including ceiling and floor effects, biased distributions, partial and incomplete representation of neurocognitive dominions, and poor calibration.<sup>23</sup> These metric properties are shared by most of the brief cognitive tests and substantially limit their potential to measure broad-spectrum cognitive function and longitudinal changes. In fact, most of the studies analysing the association between one or several brief tests and psychometric tests use analytic techniques which establish a correlation between variables but do not adequately evaluate agreement between the values obtained from 2 tests assessing the same magnitude.<sup>24</sup>

According to our study, although TYM can explain nearly half of the variance observed in mean-z, this is not enough to rationalise using it to replace an extensive neurocognitive assessment. The values obtained from TYM are not comparable to those of the neurocognitive assessment due to the large spread of z scores in these 2 tools. The problem is not only the poor calibration of TYM, especially for extreme values, but also the dispersion of its values for each mean-z. In nearly 30% of the subjects in our study, the differences between the 2 assessment tools were greater than the minimum detectable change; however, from a clinical viewpoint, the most important fact is that the direction of the deviation was unpredictable. We can conclude that, although TYM is an adequate cognitive screening tool for cognitive impairment and dementia, its use is limited. Until now, no studies have proven its sensitivity to change and its connection with such external validation criteria as neuropathology and neuroimaging findings and other biomarkers.

We may wonder why we should use a global cognitive performance index like the one applied in this study considering that some authors have observed that certain brief tests may predict functional status better than an extensive neurocognitive assessment.<sup>25</sup> The ecological validity of neuropsychological tests has been assessed; these were found to account for no more than 30% of the variance in daily life function measures.<sup>25,26</sup> We should acknowledge, nonetheless, that activities of daily living instruments have their own drawbacks<sup>27</sup>: the reliability of the data obtained from informants is highly variable and often unverifiable; measures may address a heterogeneous group of different activities (basic, instrumental, advanced, housework, social activities, hobbies, etc.) that are not uniformly included in scales;

activities of daily living may vary significantly between subjects with different lifestyles but similar sociodemographic characteristics; and a large number of tools lack the minimum metric requirements, with the result that there is a wide range of activities of daily living instruments (nearly as many as brief cognitive tests), clearly indicating that none of the existing tools is truly satisfactory.<sup>28</sup> Some of the problems inherent to ordinal scales and cognitive tests may be avoided by using modern design techniques (item response theory, Rasch analysis, etc.).<sup>27,29,30</sup> Although we accept that cognitive impairment very likely does result in functional impairment, we lack specific tools to prove this. The concept of 'mild cognitive impairment' would therefore be an artificial category derived from our own imperfect perception of reality.

From a conservative viewpoint, most healthcare professionals may agree with us that formal neurocognitive tests, which are multifaceted and relatively extensive, assess a patient's cognitive status more accurately than any brief test and, as a result, provide a more reliable evaluation of brain function, which is our ultimate purpose. In addition, their prospective validity is clearly superior to that of brief tests, precisely when accuracy is most needed: in cases of diagnostic uncertainty and when cognitive impairment seems more obvious to patients or their relatives than to clinicians.<sup>23</sup> Although a recent, thorough analysis of the ADAS-Cog has identified several significant weaknesses of this tool,<sup>31,32</sup> it is still widely used for evaluating results from clinical trials on Alzheimer disease.<sup>33–35</sup> In our opinion, the same is applicable to many other brief cognitive tests. Although the literature provides a detailed description of design requirements for brief tests,<sup>36</sup> these are not always met. A thorough neurocognitive evaluation provides accurate qualitative data, which are incomparably more valuable and complete than brief test data.

In conclusion, our results suggest that TYM scores (and possibly scores from other brief cognitive screening tests) should not be used to assess a patient's cognitive state, evaluate progression of cognitive impairment, determine changes in clinical trials, or predict the results of a more extensive psychometric test.

## Funding

This study received no public or private funding.

## Conflicts of interest

The authors have no conflicts of interest to declare.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.nrleng.2015.03.006>.

## References

1. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12:189–98.
2. Farber JF, Schmitt FA, Logue PE. Predicting intellectual level from the Mini-Mental State Examination. *J Am Geriatr Soc.* 1988;36:509–10.
3. Nakamura H, Nakanishi M, Furukawa TA, Hamanaka T, Tokudome S. Validity of brief intelligence tests for patients with Alzheimer's disease. *Psychiatry Clin Neurosci.* 2000;54:435–9.
4. Ferrero-Arias J, Turrión-Rojo MÁ. Validación de una versión española del Test Your Memory. *Neurología.* 2014. <http://dx.doi.org/10.1016/j.nrl.2013.12.009>.
5. Hughes CP, Berg L, Danziger WL, Cohen LA, Martin RL. A new clinical scale for the staging of dementia. *Br J Psychiatry.* 1982;140:566–72.
6. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology.* 1993;43:2412–4.
7. Lucas JA, Ivnik RJ, Willis FB, Ferman TJ, Smith GE, Parfitt FC, et al. Mayo's older African Americans normative studies: normative data for commonly used clinical neuropsychological measures. *Clin Neuropsychol.* 2005;19:162–83.
8. Peña-Casanova J, Blesa R, Aguilar M, Gramunt-Fombuena N, Gómez-Ansón B, Oliva R, et al. Spanish Multicenter Normative Studies (NEURONORMA Project): methods and sample characteristics. *Arch Clin Neuropsychol.* 2009;24:307–19.
9. Peña-Casanova J, Quiñones-Ubeda S, Quintana-Aparicio M, Aguilar M, Badenes D, Molinuevo JL, et al. Spanish Multicenter Normative Studies (NEURONORMA Project): norms for verbal span, visuospatial span, letter and number sequencing, trail making test, and symbol digit modalities test. *Arch Clin Neuropsychol.* 2009;24:321–41.
10. Peña-Casanova J, Quiñones-Ubeda S, Gramunt-Fombuena N, Quintana-Aparicio M, Aguilar M, Badenes D, et al. Spanish Multicenter Normative Studies (NEURONORMA Project): norms for verbal fluency tests. *Arch Clin Neuropsychol.* 2009;24:395–411.
11. Peña-Casanova J, Gramunt-Fombuena N, Quiñones-Ubeda S, Sánchez-Benavides G, Aguilar M, Badenes D, et al. Spanish Multicenter Normative Studies (NEURONORMA Project): norms for the Rey–Osterrieth complex figure (copy and memory), and free and cued selective reminding test. *Arch Clin Neuropsychol.* 2009;24:371–93.
12. Peña-Casanova J, Quiñones-Ubeda S, Gramunt-Fombuena N, Aguilar M, Casas L, Molinuevo JL, et al. Spanish Multicenter Normative Studies (NEURONORMA Project): norms for Boston naming test and token test. *Arch Clin Neuropsychol.* 2009;24:343–54.
13. Jones RN, Rudolph JL, Inouye SK, Yang FM, Fong TG, Milberg WP, et al. Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. *J Clin Exp Neuropsychol.* 2010;32:1041–9.
14. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology.* 2014;42:144–53.
15. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–10.
16. Luiz RR, Costa AJ, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. *J Clin Epidemiol.* 2003;56:963–7.
17. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol.* 1991;59:12–9.
18. Ihl R, Frölich L, Dierks T, Martin EM, Maurer K. Differential validity of psychometric tests in dementia of the Alzheimer type. *Psychiatry Res.* 1992;44:93–106.
19. Giordani B, Boivin MJ, Hall AL, Foster NL, Lehtinen SJ, Bluemlein LA, et al. The utility and generality of Mini-Mental State Examination scores in Alzheimer's disease. *Neurology.* 1990;40:1894–6.
20. Brooks JO 3rd, Yesavage JA, Taylor J, Friedman L, Tanke ED, Luby V, et al. Cognitive decline in Alzheimer's disease: elaborating on the nature of the longitudinal factor structure of the Mini-Mental State Examination. *Int Psychogeriatr.* 1993;5:135–46.
21. Dick JP, Guiloff RJ, Stewart A, Blackstock J, Bielawska C, Paul EA, et al. Mini-mental state examination in neurological patients. *J Neurol Neurosurg Psychiatry.* 1984;47:496–9.
22. Sugarman MA, Axelrod BN. Utility of the Montreal cognitive assessment and mini-mental state examination in predicting general intellectual abilities. *Cogn Behav Neurol.* 2014;27:148–54.
23. Wouters H, Appels B, van der Flier WM, van Campen J, Klein M, Zwinderman AH, et al. Improving the accuracy and precision of cognitive testing in mild dementia. *J Int Neuropsychol Soc.* 2012;18:314–22.
24. Bland JM, Altman DG. Agreed statistics: measurement method comparison. *Anesthesiology.* 2012;116:182–5.
25. Royall DR, Lauterbach EC, Kaufer D, Malloy P, Coburn KL, Black KJ, et al. The cognitive correlates of functional status: a review from the Committee on Research of the American Neuropsychiatric Association. *J Neuropsychiatry Clin Neurosci.* 2007;19:249–65.
26. Van der Elst W, van Boxtel MP, van Breukelen GJ, Jolles J. A large-scale cross-sectional and longitudinal study into the ecological validity of neuropsychological test measures in neurologically intact people. *Arch Clin Neuropsychol.* 2008;23:787–800.
27. Lindeboom R, Vermeulen M, Holman R, de Haan RJ. Activities of daily living instruments: optimizing scales for neurologic assessments. *Neurology.* 2003;60:738–42.
28. Fieo RA, Austin EJ, Starr JM, Deary IJ. Calibrating ADL–IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC Geriatr.* 2011;11:42.
29. McGrory S, Doherty JM, Austin EJ, Starr JM, Shenkin SD. Item response theory analysis of cognitive tests in people with dementia: a systematic review. *BMC Psychiatry.* 2014;14:47.
30. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated calibrating cognitive tests and reduced bias in estimated rates of decline. *J Clin Epidemiol.* 2008;61:1018–27.
31. Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, et al. Putting the Alzheimer's cognitive test to the test I: traditional psychometric methods. *Alzheimers Dement.* 2013;9 Suppl. 1:54–9.
32. Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, et al. Putting the Alzheimer's cognitive test to the test II: Rasch measurement theory. *Alzheimers Dement.* 2013;9 Suppl. 1: S10–20.
33. Salloway S, Sperling R, Fox NC, Blennow K, Klunk W, Raskind M, et al. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N Engl J Med.* 2014;370:322–33.

34. Doody RS, Raman R, Farlow M, Iwatsubo T, Vellas B, Joffe S, et al. A phase 3 trial of semagacestat for treatment of Alzheimer's disease. *N Engl J Med.* 2013;369:341–50.
35. Doody RS, Thomas RG, Farlow M, Iwatsubo T, Vellas B, Joffe S, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med.* 2014;370:311–21.
36. Peña-Casanova J, Monllau A, Gramunt Fombuena N. La psicometría de las demencias a debate. *Neurología.* 2007;22:301–11.